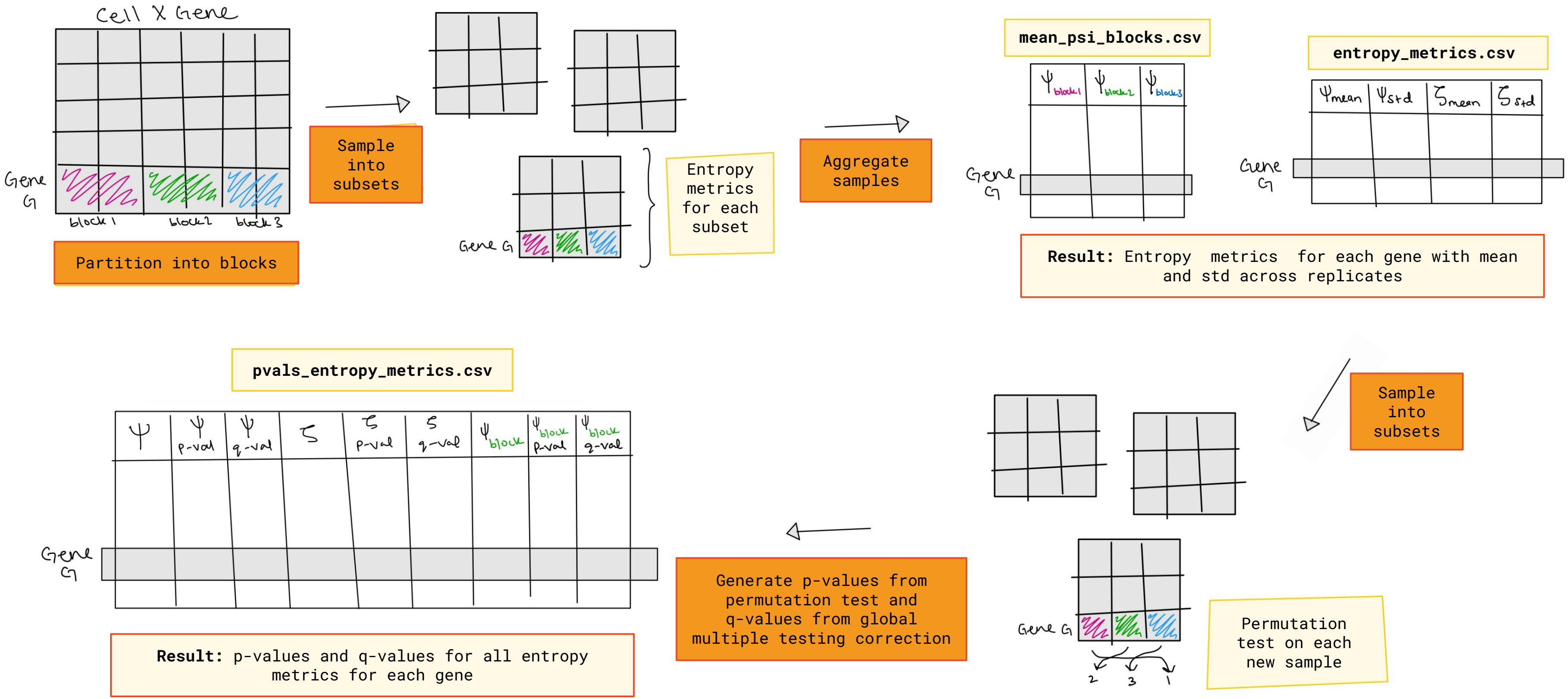
ember entropy metrics for biological exploration

Nikhila P. Swarna
October 9 2025

Pachter Lab
Division of Biology and Bioengineering
California Institute of Technology

Download link
https://github.com/pachterlab/ember

light_ember workflow



Partition into blocks

Sample into subsets

Entropy metrics for each subset

Aggregate samples

mean_psi_blocks.csv

ψ_{block1}	ψ_{block2}	ψ_{block3}

entropy_metrics.csv

ψ_{mean}	ψ_{std}	ζ_{mean}	ζ_{std}

Result: Entropy metrics for each gene with mean and std across replicates

Sample into subsets

Permutation test on each new sample

Generate p-values from permutation test and q-values from global multiple testing correction

pvals_entropy_metrics.csv

ψ	ψ p-val	ψ q-val	ζ	ζ p-val	ζ q-val	ψ_{block}	ψ_{block} p-val	ψ_{block} q-val

Result: p-values and q-values for all entropy metrics for each gene

ember **e**ntropy **m**etrics **f**or **b**iological **e**xplo**r**ation

light_ember - one stop shop for generating entropy metrics and p-values

INPUT

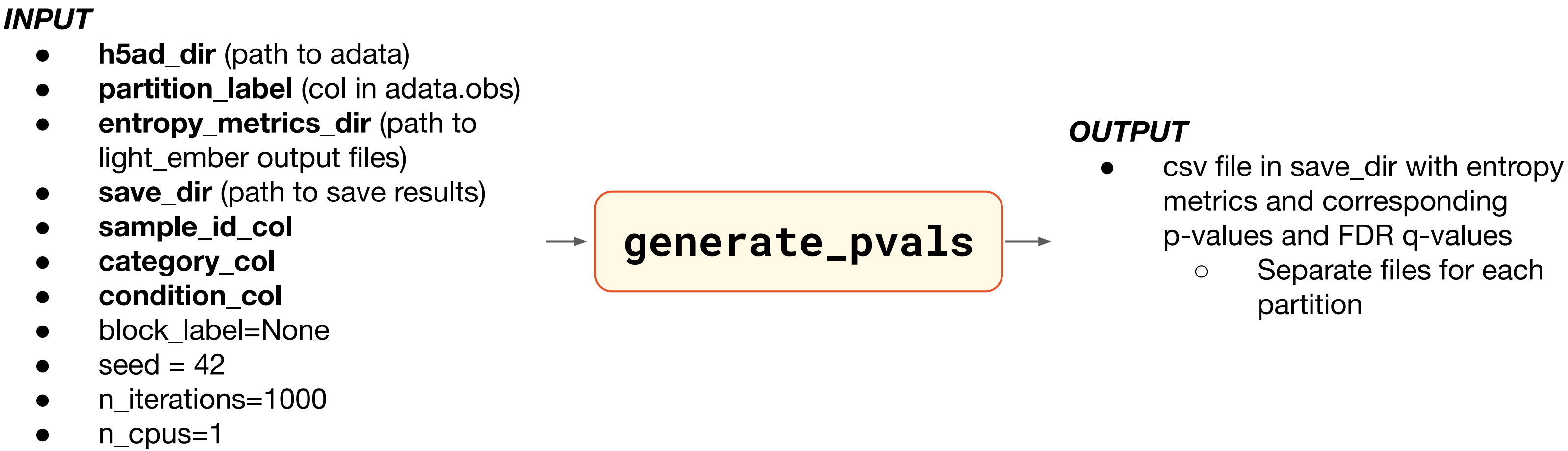
- **h5ad_dir** (path to adata)
- **partition_label** (col in adata.obs)
- **save_dir** (path to save results)
- sampling= True
  - sample_id_col=None
  - category_col=None
  - condition_col=None
  - num_draws=100
  - save_draws = False
  - seed = 42
- partition_pvals=True
- block_pvals=False
  - block_label=None
  - n_pval_iterations=1000
- n_cpus=1 (for parallel processing of sampling)

OUTPUT

- csv file in save_dir with all entropy metrics
- csv file in Psi_block_df folder with psi block
  - Separate file for pvals
  - Separate files for each partition
  - Alternate file names depending on sampling on or off.

ember entropy metrics for biological exploration

generate_pvals - manual access to generate pvals after initial investigation using **light_ember**



INPUT

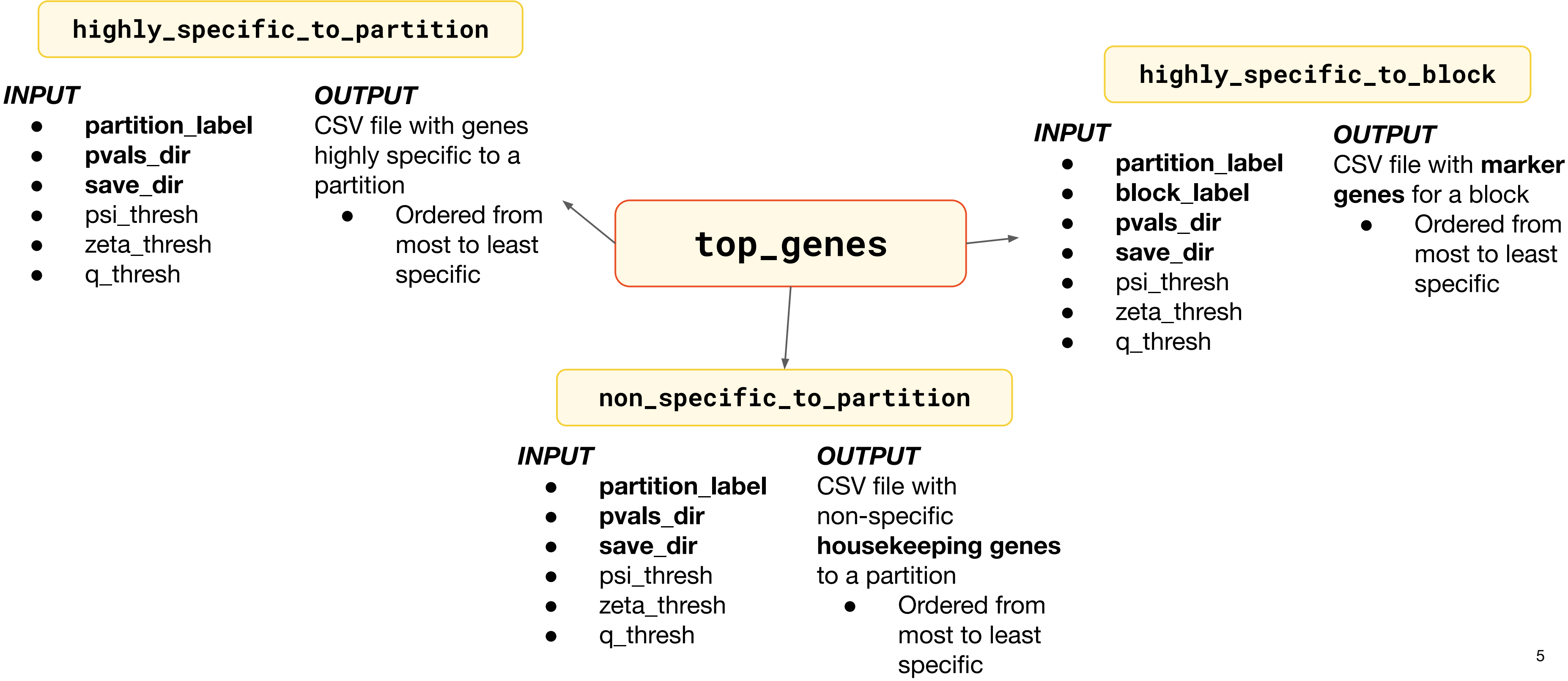
- **h5ad_dir** (path to adata)
- **partition_label** (col in adata.obs)
- **entropy_metrics_dir** (path to light_ember output files)
- **save_dir** (path to save results)
- **sample_id_col**
- **category_col**
- **condition_col**
- block_label=None
- seed = 42
- n_iterations=1000
- n_cpus=1

generate_pvals

OUTPUT

- csv file in save_dir with entropy metrics and corresponding p-values and FDR q-values
  - Separate files for each partition

ember **e**ntropy **m**etrics for **b**iological **e**xplo**r**ation



top_genes

highly_specific_to_partition

INPUT

- **partition_label**
- **pvals_dir**
- **save_dir**
- psi_thresh
- zeta_thresh
- q_thresh

OUTPUT

CSV file with genes highly specific to a partition

- Ordered from most to least specific

highly_specific_to_block

INPUT

- **partition_label**
- **block_label**
- **pvals_dir**
- **save_dir**
- psi_thresh
- zeta_thresh
- q_thresh

OUTPUT

CSV file with **marker genes** for a block

- Ordered from most to least specific

non_specific_to_partition

INPUT

- **partition_label**
- **pvals_dir**
- **save_dir**
- psi_thresh
- zeta_thresh
- q_thresh

OUTPUT

CSV file with non-specific **housekeeping genes** to a partition

- Ordered from most to least specific

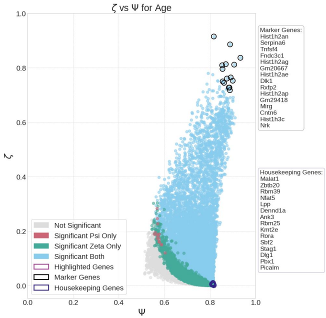
ember **e**ntropy **m**etrics for **b**iological **e**xplo**r**ation

plots

`plot_partition_specificity`

INPUT

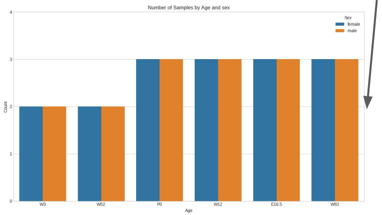
- **partition_label**
- **pvals_dir**
- **save_dir**
- highlight_genes
- fontsize
- color_palette



`plot_sample_counts`

INPUT

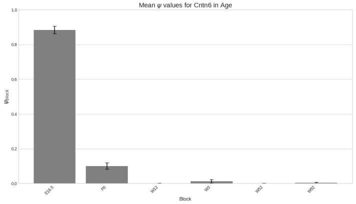
- **h5ad_dir**
- **save_dir**
- **sample_id_col**
- **category_col**
- **condition_col**
- fontsize



`plot_psi_blocks`

INPUT

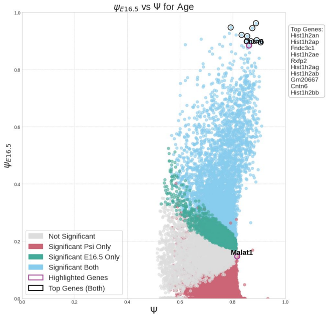
- **gene_name**
- **partition_label**
- **psi_block_df_dir**
- **save_dir**
- fontsize



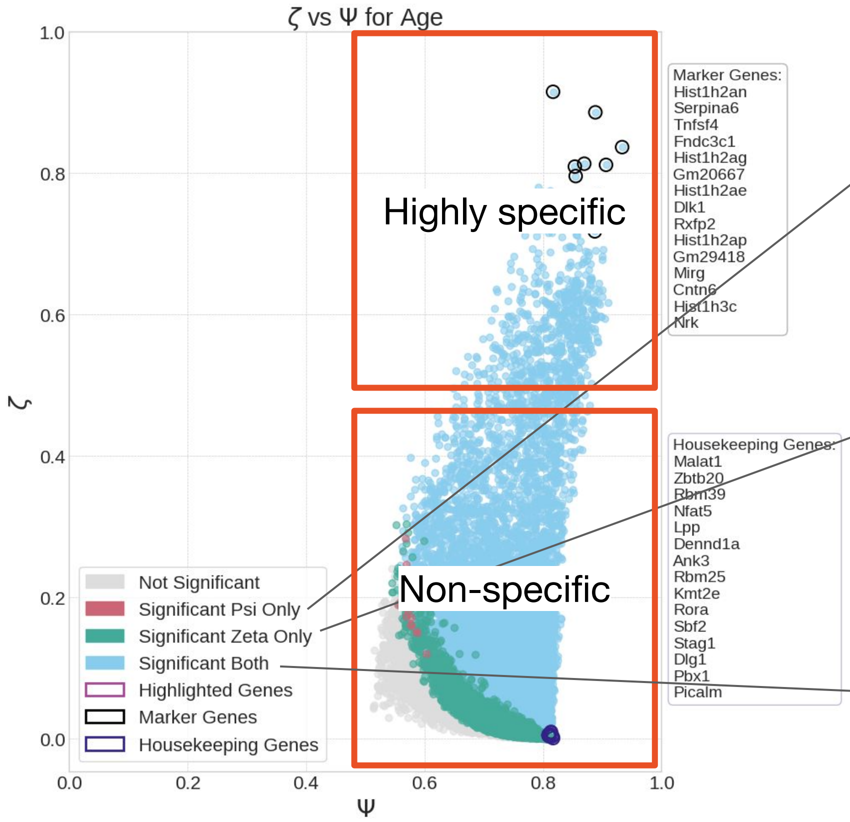
`plot_block_specificity`

INPUT

- **partition_label**
- **block_label**
- **pvals_dir**
- **save_dir**
- highlight_genes
- fontsize
- color_palette



How to interpret **`plot_partition_specificity`**

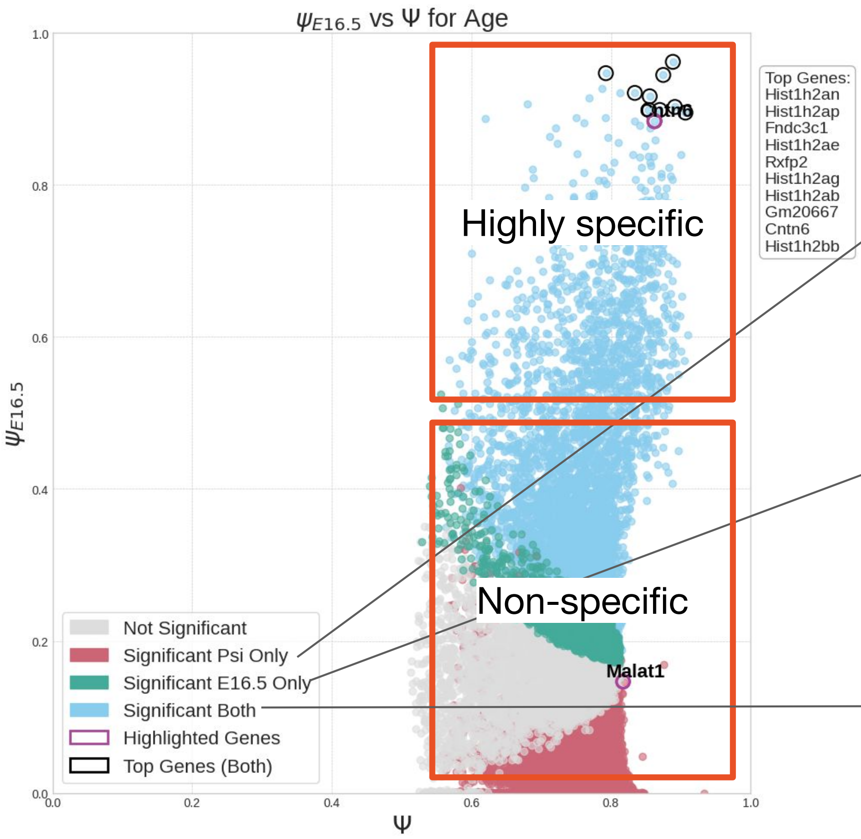


Significant by Psi only - Genes with unreliable Zeta score after permutation testing, **cannot reliably infer specificity**

Significant by Psi only - Genes with unreliable Psi score after permutation testing, **cannot reliably infer specificity**

Significant by Both - Genes with reliable Zeta and Psi scores after permutation testing with **reliable specificity scores**!

How to interpret **`plot_block_specificity`**



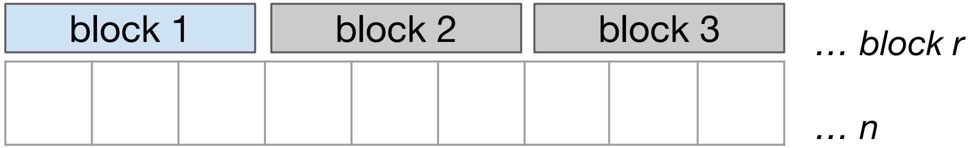
Significant by Psi only - Genes with unreliable psi_block score after permutation testing = **cannot reliably infer specificity**

Significant by psi_block only - Genes with unreliable Psi score after permutation testing = **cannot reliably infer specificity**

Significant by Both - Genes with reliable psi_block and Psi scores after permutation testing = **reliable specificity scores**!

Defining entropy metrics for biological exploration

For a given gene in a count matrix that can be partitioned into r blocks (based on sex, strain, cell type, tissue, etc), we introduce **3 measures of specificity**:



- Psi (Ψ)
- Psi$_{\text{block}}$ (Ψ_{block})
- Zeta (ζ)

Ψ	Ψ_{block}	ζ
Fraction of information explained by partitioning	Specificity to a block	Specificity to a partition

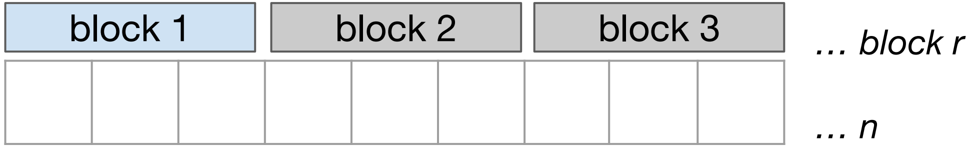
Ψ- Information Fraction by Partition

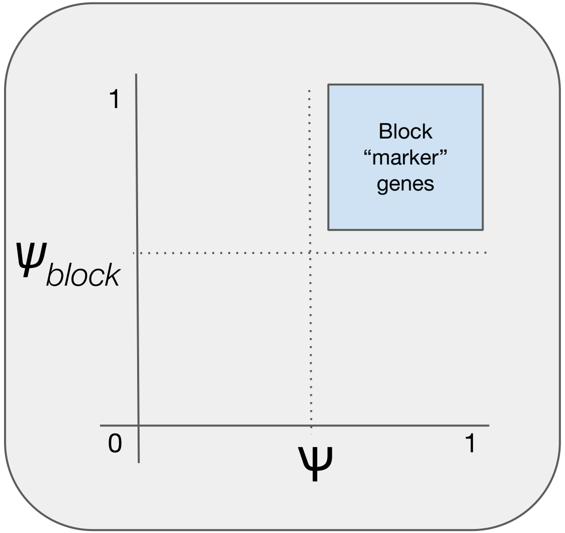
The fraction of information explained by using a particular partition on gene *g*'s counts is given by:

$$\Psi = \frac{E_W}{E_T} = 1 - \frac{E_B}{E_T}$$

The Specificity of Information to Block, denoted by ψ_{block}, is the contribution of each block to Ψ:

$$1 = \boxed{\begin{matrix}\psi_1 \\ \frac{p_1 E_1}{E_W}\end{matrix}} + \boxed{\begin{matrix}\psi_2 \\ \frac{p_2 E_2}{E_W}\end{matrix}} + \cdots + \boxed{\begin{matrix}\psi_r \\ \frac{p_r E_r}{E_W}\end{matrix}}$$



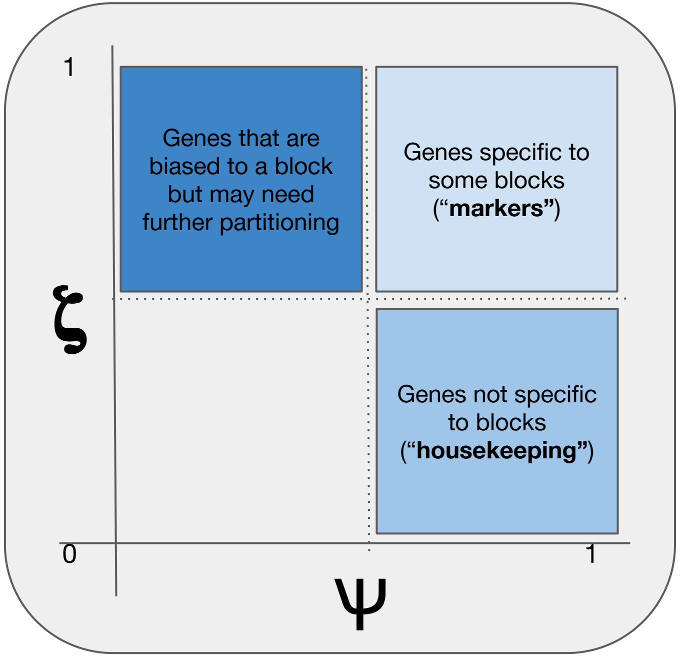


ζ - Specificity of Information to Partition

The specificity of information to a partition (ζ) is given by:

$$\zeta = 1 - \frac{H(\psi_{\text{blocks}})}{\log r}$$

Comparison of the SIB distribution to the uniform distribution



Genes that are biased to a block but may need further partitioning

Genes specific to some blocks (“**markers”**)

Genes not specific to blocks (“**housekeeping”**)

How to select category and condition

Category - Mouse strain
Condition - Sex

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Mouse strain	Strain A				Strain B				Strain C				Strain D			
Sex	Male		Female		Male		Female		Male		Female		Male		Female	

Number of unique draws
= (Number of replicates per category-condition group)^(Number of category-condition groups)
= 2^8 = **256**

One example draw:

Sample	1	3	5	7	9	11	13	15
Mouse strain	Strain A		Strain B		Strain C		Strain D	
Sex	Male	Female	Male	Female	Male	Female	Male	Female

How to select category and condition

Category - Cell line
Condition - Gene perturbation

Sample	1	2	3	4	5	6	7	8	9	10	11	12
Cell line	Cell line A						Cell line B					
Gene perturbation	Wildtype		Overexpression		Knockout		Wildtype		Overexpression		Knockout	

Number of unique draws
= (Number of replicates per category-condition group)^(Number of category-condition groups)
= 2^6 = **64**

One example draw:

Sample	1	3	5	7	9	11
Cell line	Cell line A			Cell line B		
Gene perturbation	WT	OE	KO	WT	OE	KO

How to select category and condition

Category - Mouse strain
Condition - Sex

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Mouse strain	Strain A				Strain B				Strain C					Strain D		
Sex	Male	Female			Male		Female		Male			Female		Male		Female

Number of unique draws
= ∏ (Number of replicates per category-condition group)
= 1*3*2*2*3*2*2*1 = **144**

One example draw:

Sample	1	2	5	7	9	12	14	16
Mouse strain	Strain A		Strain B		Strain C		Strain D	
Sex	Male	Female	Male	Female	Male	Female	Male	Female

Note: #1 and #16 will appear in every draw since there are no replicates for these groups